Homework 3: Variable Elimination, Gibbs Sampling, and Value of Information

Release: October 14, 2014
Due: October 23, 2014, 11:59pm

1 Programming Component [30 points]

(30 points) Refer to the course website for links and description of the programming component, at http://www.cs.cmu.edu/~15381.

2 Written Component [30 points]

To receive credit for your work, upload a pdf containing your solutions to AutoLab under “writeups” at https://autolab.cs.cmu.edu.

Problem 1 – Variable Elimination [6 points]

Consider the example from the slide 81 of the lecture on October 2, 2014 (here). We are interested in Pr(B). First consider variable elimination, eliminating variables in the order A, S₁, ..., S₉. What is the size of the largest factor, in terms of the number of variables, in this elimination process? Now, consider variable elimination, eliminating variables in the order S₁, ..., S₉, A. What is the size of the largest factor in this case?

Problem 2 – Value of Information [14 points]

Consider a medical diagnosis case where a patient potentially has a specific disease. Let Y be a random variable that indicates whether a patient has the disease (Y = 1) or not (Y = 0). We wish to choose a course of action for a patient. We denote the set of possible actions by A. A patient’s welfare depends on whether or not the patient has the disease, and the action chosen by us. For an action a ∈ A and health status of patient Y = y, we denote the welfare of the patient by U(a, y) (larger numbers indicate better outcomes). Since we cannot observe Y directly, we base our decision on n medical tests whose results are given by random
variables $X_1, \ldots, X_n$. We must determine which set of tests to perform in order to obtain a good outcome for the patient.

For a set $V \subseteq \{1, \ldots, n\}$, we use $\bar{X}_V = \{X_i | i \in V\}$ to indicate the set of random variables that correspond to tests in $V$ and we denote their joint probability by $\Pr(\bar{X}_V)$. We also use $\bar{x}_V$ to indicate the observed value of $\bar{X}_V$.

One way to evaluate a set of tests is their value of information,

$$V_{OI}(V) = \sum_{\bar{x}_V} \Pr(\bar{x}_V) \max_{a \in A} \sum_{y \in \{0,1\}} \Pr(y | \bar{x}_V) U(a, y),$$

which captures the expected utility that results from choosing an action based on the outcome of the tests $\bar{X}_V$.

As a simple example, let the probability of a patient having the disease be $\Pr(Y = 1) = 0.5$. Let there be 3 possible actions whose payoffs are given by the table below:

<table>
<thead>
<tr>
<th>$U(y, a)$</th>
<th>$A = 1$</th>
<th>$A = 2$</th>
<th>$A = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>1</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>-5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Assume that there are 2 possible tests, $X_1$ and $X_2$. For either test, value of 1 indicates that the patient is tested positive for the disease and 0 indicates that the patient is tested negative for the disease. Unfortunately, our tests are not entirely accurate. For each test, $\frac{1}{2}$ of the time the test returns the health status of the patient ($X = Y$) and $\frac{1}{2}$ of the time it returns a value chosen uniformly at random from $\{0, 1\}$. Furthermore, the tests are conditionally independent given $Y$.

1. (5 points) Fill out the value of the following probabilities.

$$\Pr(Y = 1 | X_1 = 1) =$$
$$\Pr(Y = 1 | X_1 = 0) =$$
$$\Pr(Y = 1 | X_1 = 1, X_2 = 1) =$$
$$\Pr(Y = 1 | X_1 = 0, X_2 = 1) =$$
$$\Pr(Y = 1 | X_1 = 1, X_2 = 0) =$$
$$\Pr(Y = 1 | X_1 = 0, X_2 = 0) =$$

$$\Pr(X_1 = 1) =$$
$$\Pr(X_1 = 1, X_2 = 1) =$$
$$\Pr(X_1 = 1, X_2 = 0) =$$
$$\Pr(X_1 = 0, X_2 = 1) =$$
$$\Pr(X_1 = 0, X_2 = 0) =$$

2. (6 points) Compute the value of $V_{OI}(\{1\})$, $V_{OI}(\{2\})$ and $V_{OI}(\{1, 2\})$. Show your work.

3. (3 points) For a given set $S$, a function $f : 2^S \to \mathbb{R}$ is submodular if for any $V, W \subseteq X$ such that $V \subseteq W$ and for any $s \in S \setminus W$ we have that

$$f(V \cup \{s\}) - f(V) \geq f(W \cup \{s\}) - f(W).$$

Show that the above example implies that Value of Information is in general not a submodular function.

2
Problem 3 – Gibbs sampling and its limitations [10 points]

Gibbs sampling (and MCMC algorithms in general) is an extremely useful tool for scenarios where it is infeasible to sample from the joint distribution. However, as with most heuristic algorithms, Gibbs sampling also has its limitations. In this homework we will explore these limitations.

Getting stuck

Consider a distribution over the four states
\[ p(x = 0, y = 0) = 0.5, p(x = 0, y = 1) = 0, p(x = 1, y = 0) = 0, p(x = 1, y = 1) = 0.5 \]

a) (2 points) Fill out the following conditional probabilities:
\[ p(x = 1|y = 0) = ? \]
\[ p(x = 1|y = 1) = ? \]

b) (2 points) Given starting state \((x = 1, y = 1)\), write the probability distribution over the states returned from Gibbs sampling, and give a brief explanation of why Gibbs sampling behaves the way it does here.

Convergence issues

Now consider the set of Boolean vectors of length 100:
\[ \{ x : \forall i = 1, \ldots, 100 : x_i \in \{0, 1\} \} \]

with the all-zero vector having probability \( p(x = 0) = 0.5 \), and \( p(x) = \frac{1}{2^{100}-1} \) for all other \( x \). Consider the Gibbs sampling algorithm presented on slide 46 (pages 536 – 538 in Russel and Norvig may also be useful reading):

http://www.cs.cmu.edu/~ebrun/381_approx_bn2.pdf

where we iterate over all the variables from 1 to 100, and sample a value for each variable in turn.

c) (2 points) Given initial state \( x_i = 0 \) for all \( i \), what is the probability of sampling 1 at step \( j \) of Gibbs sampling, assuming that you sampled 0 at all \( k < j \) at this iteration. You don’t have to compute actual values, but we should be able to plug your solution directly into a calculator.

d) (2 points) Now compute the probability of updating to any state where \( x_i = 1 \) for one or more \( i \). You don’t have to compute actual values, but we should be able to plug your solution directly into a calculator.
(2 points) Briefly explain why this shows that convergence to the actual distribution will take a large amount of time in expectation. As a sidenote, it can be shown that it is similarly improbable to reach the all-zero state from any other state, but this requires somewhat more tricky probability calculations.